The International Journal of Digital Curation
Issue 1, Volume 3 | 2008

Curation of Laboratory Experimental Data as Part of the Overall Data Lifecycle

Jeremy Frey,
Professor of Physical Chemistry,
University of Southampton

June 2008

Abstract

The explosion in the production of scientific data in recent years is placing strains upon conventional systems supporting integration, analysis, interpretation and dissemination of data and thus constraining the whole scientific process. Support for handling large quantities of diverse information can be provided by e-Science methodologies and the cyber-infrastructure that enables collaborative handling of such data. Regard needs to be taken of the whole process involved in scientific discovery. This includes the consideration of the requirements of the users and consumers further down the information chain and what they might ideally prefer to impose on the generators of those data. As the degree of digital capture in the laboratory increases, it is possible to improve the automatic acquisition of the 'context of the data' as well as the data themselves. This process provides an opportunity for the data creators to ensure that many of the problems they often encounter in later stages are avoided. We wish to elevate curation to an operation to be considered by the laboratory scientist as part of good laboratory practice, not a procedure of concern merely to the few specialising in archival processes. Designing curation into experiments is an effective solution to the provision of high-quality metadata that leads to better, more re-usable data and to better science.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

In the traditional approach to scientific discovery, practising physical scientists are taught to take great care in the recording of their observations. In principle they are also taught to think about the hypothesis, the reasons for the investigation, and thus plan the experiments in advance. However, even when this planning is replaced by more *ad hoc* discovery techniques, the need for detailed recording of what was observed, in its context, is an essential part of the scientific process, albeit one not always fully honoured in the practice. The explosion in the rate of production of scientific data due to new technology and automation is placing the whole system of scientific investigation and publication under considerable strain (Allen, 2004; Dennis, 2002; Patterson, 2003; Russo, 2002).

Provenance and Dissemination

In undertaking laboratory experiments scientists attempt to record the origin of all the materials, the details of the equipment used, and the design and results of the experiments undertaken. They expect to be able to trace the possible influences on their experiments so that as they successively re-design and elaborate their investigations, they narrow down the possible causes of the phenomena under investigation. They are therefore often acutely aware of the importance of the provenance of the materials and data employed in their investigations. Since they hope their results, once refined, will be used by others, it behoves them to consider how those subsequent users will verify and exploit the information they have published. Historically, this concern with the output of the investigations has not been explicit while the investigation is proceeding, but has mainly come into play once the investigation is finished and the work "written up". The concept of publication@source attempts to change this perspective to one in which the issues of re-use of data are considered at the time the information is created (Frey, De Roure, & Carr, 2002).

Implicit of course is the need to preserve the data carefully. Can students find and accurately identify the correct spectrum of a sample taken several years earlier at the start of their research work? A final year PhD student is often heard to remark, "I wish I had taken as careful notes at the start as I do now, it would be so much easier to write up". Can collaborators be sure they have access to (all) the information they need? Will they still be able to access this information in the future? Indeed the growing scope of collaborative research, on a global scale, is one of the driving forces for the improved description of information and e-science-facilitated interactions (Frey et al., 2006).

Good Laboratory Practice

In most cases these issues have been covered under what would be called "good laboratory practice". Proper record keeping is an essential part of all high-quality research. A record of what is planned and what outcomes were observed is a necessary part of this process. In the past, records of, for example, spectra would be displayed on paper. That was how a chart recorder generated them; no computer was involved. The paper record would be glued into the laboratory notebook (the 'cut-and - paste' technique), so that the association between the *notes* and the *spectra* was

maintained. With increasingly complex instrumentation and data analysis with very large data files, it is no longer feasible to attempt a 'cut-and-paste' operation; the data simply cannot be represented on paper.

Even if 'cut and paste' could be used, printing transforms a computer-readable data file into a form that cannot easily be re-used. To re-use the figure in subsequent processing, the data would have to be re-typed into computer-readable form. So 'cut and paste' is an example of a preservation technique that is very good at maintaining the link between the materials – experimental investigation – the picture of the spectrum – the notes; but with which significant functionality is lost (Figure 1).

Figure 1. The laboratory notebook needs to be in the laboratory where it is exposed to many dangers. The blue carbon copy is one method of ensuring preservation as long as the pages are removed from the laboratory notebook. (Picture courtesy of the Smart Tea Group.)

In practice the spectral and similar data are now held in a data file, on a computer hard disk, which may be backed up to a CD or DVD (perhaps more than one, in case of failures). The file name, together with the directory, is recorded in the laboratory notebook, with the most assiduous students providing archived details and an index. A formal structure driven by Health and Safety (e.g. COSHH – Control of Substances Hazardous to Health[1]) does provide a driving force to support the correct process. Similarly a standardised directory structure for the data files of the form, *<researcher>/<data>/<number>/* allows for some checking and retrieval of "lost data" by correlating file names with the dates recorded in the laboratory notebook. During MPhil/PhD transfer vivas at the School of Chemistry at Southampton, the students' laboratory notebook is routinely checked for inadequate entries; the less assiduous students are frequently asked whether their supervisor could find their file without those students' personal intervention. A rare few students have indexed their laboratory notebooks.

This practice has been explained in some detail to underline the concerns that are already present in existing good laboratory investigations. Many, if not all, of the same concerns apply to fieldwork and other types of study. What is at issue now is:

[1] COSHH http://www.hse.gov.uk/coshh/

- how these practices may fall short of what is required to support the current and future scale of scientific investigation, and;
- how new types of investigations may become possible with improvements in the recording of experiments.

The explosion in the amount of data has driven a great deal of the current progress in these areas. It has always been clear that good planning leads to improved execution of experiments. Digital methods do not change this; they simply make the utility of plans even more evident as they are applied to ever-increasing amounts of data. Digital plans are more readily transferred, shared and adapted and if appropriately written can be used as the basis of automation in the laboratory.

It has always been clear that good planning leads to good execution of experiments. Digital methods and the explosion in the amount of data do not change this; they simply make the value and utility of these plans even more evident.

The Data Explosion

The data explosion in scientific research has had a major impact on biology and increasingly on chemistry (Allen, 2004; Dennis, 2002; Houlton, 2001; Patterson, 2003; Russo, 2002). The rapidly expanding capability to generate data, to make samples and analyse them in a high-throughput manner, and the increasing numbers of practitioners involved in global scientific research, means there is greater difficulty in ensuring data and results can be accessed, understood and used. Creating documentation, and keeping it up to date and accurate is a challenging and time-consuming task; a task that can take significantly longer than producing and characterising some materials. This is a phenomenon known to all users of computer programs who frequently find that documentation either does not exist or relates to an older version of the program. Computer scientists who maintain that 'the program documents itself' have their parallel in the laboratory, where the mantra is usually, "well *I know* which sample is which".

Curation: Who Is Responsible? Who Should Be Responsible?

When the scientific process has reached a certain stage, researchers write a paper, which is then, if sufficiently appreciated by referees, published and enters the formal scientific record. Data may be contained within the paper or as part of some supplementary material. Most researchers' concerns stop there, other than to see if others cite their work, something of increasing importance in deriving recognition and reward. The quality of the journal is important and it may occur to researchers that some widely regarded journals will be held by more libraries round the world than less prestigious ones. They are however unlikely to spend time worrying about the access to that journal in 10, 20 or 100 years time; that is perceived as the concern of the library community. The publishing and library communities have responded extremely well to the need to make their collections available in new ways (e.g. a pdf file delivered over the Web, HTML pages, and the ability to search and locate information). In many cases they have taken on board the resulting drive to publish different material (audio, video, raw data files, databases etc.) alongside the more usual text.

The true value of the research over time will inevitably be enhanced if the work is accessible, the data re-useable, the explanations fully supported. Access to raw data provides additional provenance, and the ability to check and re-work material in the light of new evidence or theory (e.g. meta-studies, etc).

Investigators should be more concerned with the way their outputs are presented, how they are held and what readers can (and are permitted to) do with them. Often the focus has been only on the journal paper and not on the underlying data. The journal paper may very adequately present the main intellectual ideas, but often does a poorer job at presenting all the materials that are needed to justify the conclusions. Even if tables of data are provided, without context they can engender misinterpretation and ambiguity. This becomes increasingly important, as researchers need to integrate data from a wide range of sources.

In the past, data were collated into validated collections (e.g. thermodynamic data, kinetics data, etc., see section on Validation) and these collections were published as documents. Assembling this information took time and effort from dedicated teams. In many cases the government funding is no longer available and current databases of such information need to be developed in a much more automated manner. This means the original source for the data is even more important than in the past. It also means that data need to be discoverable, in context, without human intervention. If, as authors, we fail to make the data apparent, we lose an opportunity to expose our work.

If the correct context for the data contained within a publication is not provided at source, it is often a complex, expensive and error-ridden process to try to re-derive it at a later stage. Well-organised, semantically rich, digital data capture can help to avoid this situation and can also provide clear short- and long-term benefits. It will immediately facilitate sharing and collaboration between researchers, especially in interdisciplinary teams, as well as having direct benefits in organising local short-term storage and archiving as well as longer-term preservation. All the needs of curation will be served, but in a way that has an obvious benefit for the creators as well as subsequent users (who are often the same people). In the future if the material is not discoverable by automated computer means, without significant human intervention it could, to all intents and purposes, be buried.

Organisation of This Paper

Following this scene-setting, the paper will report experiences grappling with these issues in the face of the data explosion, based on work in a series of projects including CombeChem, Smart Tea, R4L, e-Bank and e-Crystals. The paper will cover an initial electronic implementation of the Laboratory Notebook, showing how it can capture much of the smaller data and metadata that emerge from laboratory work. It will explore our use of repositories in the laboratory, showing how the ability to keep more data from intervening analysis stages can provide important provenance information to compensate for the lack of authoritative validation. Extensions to traditional methods of dissemination are discussed. Finally the advantages of capturing experimental data in semantic blogs are shown, allowing publication from the source. The conclusions demonstrate how this set of capabilities brings curation back where it should be, firmly in the hands of the laboratory scientist.

Electronic Laboratory Notebooks: CombeChem and the SmartTea Project

The CombeChem programme of work[2], which was initially funded under the UK e-Science Programme, takes a much more holistic view of the *creation – analysis – publication – use/re-use* cycle for scientific (and more specifically chemical) data and information. It recognised the central roles of the small-scale researcher creating and interacting with data. It was (one of) the first e-Science projects to promote strongly a data-centric view of e-Science together with the importance of understanding and developing high-quality and efficient human-computer interfaces if the promises of e-Science were to be achieved.

One example of how the digital world can improve the capture of information at source and aid the curation process is the increasing use being made of electronic laboratory notebooks (ELNs). These assume many forms, but in the CombeChem Programme and Smart Tea Project[3] the ELN was considered to be a replacement and improvement for the traditional notebook as the essential companion for the scientist *in the laboratory*, i.e. the experimental scientist's record and guide.

Dartmouth College formally advises its students on the importance of the laboratory notebook and the requirements for accurate note taking. It is considered to be the permanent, documented and primary record of laboratory observations[4]. All these aspects are essential to the proper progress of the scientific enquiry. The documentation asserts that observations should never be collected on notepads, filter paper or other temporary paper for later transfer into a notebook. The warning that "If you are caught using the 'scrap of paper technique', your improperly recorded data may be confiscated by your TA", highlights the importance of this approach. From a perspective further down the data chain these requirements can be seen as the necessary steps to ensure the correct provenance of the information recorded and can be considered as an important pre-requisite for proper curation. The springboard for the Smart Tea ELN was the realisation that the chemists had to record a plan for their experiments in advance for safety reasons. Capturing this record in a digital form not only starts the curation process one stage back from the actual experiment, but also provides a framework (i.e. initial metadata) on which to hang the experimental observations, even if, as is sometimes the case, the experiment turns out rather differently from the anticipated plan.

Planning and Safety

An essential component of sound experimental procedure is planning. The need to plan has been re-enforced in chemistry by the requirements of Health and Safety legislation, and in particular the COSHH, which requires a detailed plan to be produced prior to undertaking experiments with potentially hazardous materials in order to minimise risk. This plan can then be used to provide a digital framework to support the subsequent experiment and ensure that the acquired metadata are of high quality. It is essential to ensure that the researchers consider themselves part of this agenda and that the capture of the information occurs in a manner that is easy, responsive and imposes no additional steps. Indeed, it is advisable that the plan be

[2] Combechem http://www.combechem.org/
[3] The Smart Tea Project http://www.smarttea.org/
[4] Dartmouth College: ChemLab: The Chemistry 3/5 & 6 Laboratories http://www.dartmouth.edu/~chemlab/

presented as clearly advantageous in the short term, even if it is only over the longer term that the advantages actually accrue. To achieve this end requires a major commitment not only to user-centred design but also to all aspects of usability of software (see Figure 2).

Figure 2. An example of a tablet PC as an interface onto the virtual electronic laboratory notebook (from the Smart Tea Group).

In the CombeChem Project this was achieved using a process of "design by analogy", which is reported in the SmartTea discussions (Frey et al., 2003; Hughes et al., 2004; schraefel et al., 2004). The Smart Tea Project successfully demonstrated that it was possible to create an environment in which it was easy for the chemists, the software engineers and the computer scientists to work together on the same "experiment".

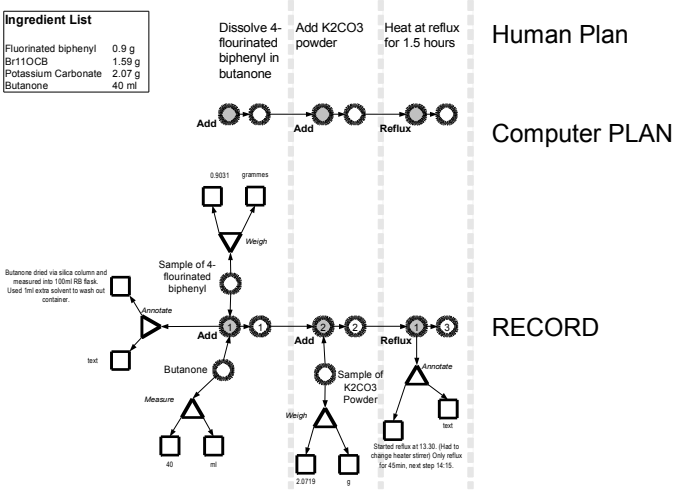

Figure 3. The plan and process record of the Smart Tea experiment. The human plan resembles a "to do" list, the computer plan makes clear the relationship between performing a process (i.e. mixing) and the result of that process. The record contains all the details of how much material was added.

This experiment was "making a cup of tea". Making tea provided a rich enough example to illustrate the essential details of chemical synthetic laboratory work in a way that allowed all researchers to question and understand not only why something was done but also what needed to be done subsequently. It led directly to software that was user-friendly and enabled the capture of high-quality metadata about the experiments (Figure 2).

Semantic Web and the ELN

There are now many ELNs available commercially but these are mostly built on conventional document management systems and relational databases[5]; a market survey is available from Atrium Research (2006). A review from a general scientific perspective has been published in *Nature* (Butler, 2006). The Southampton ELN is set apart from the mainstream in two ways: firstly it is intended that the ELN be accessed in many ways, one of which enables recording data in the laboratory as the experiments proceed, for example by using the tablet PC but the ELN exists on the network and is not "in" the tablet. The second is in the nature of the record that is captured (Figure 3).

It is the capture of the relationships, the stuff of the semantic web that set this aspect of the CombeChem agenda apart from other work in these areas. For example, in the capture of a chemical synthesis, the method is captured as a series of linked steps, for which the materials, and processes and the way they are linked are clearly stated and explicitly recorded in RDF; not just as simple free text. This allows a much greater degree of inference and processing of the experimental description to be undertaken. For example, it is relatively easy to search a conventional ELN for materials used in a process that may be incompatible (by comparison against a list of incompatible reagents); but to know whether they are a potential problem for the method requires the awareness that they are used in the same step. The linking of steps means that we can also highlight cases in which the result of one step needs to be carefully cleaned before the next step because of potential incompatibilities.

Using the ELN

Several research students in Southampton have used the ELN. Some initial difficulties with the software slowed down its adoption; however the experience of those who have used it is very positive. Even the students who have still kept a paper notebook in parallel have found the ELN very useful. As well as the expected benefits of clarity of record and safe backup of data, the additional detail needed to provide a full record of the steps in a synthetic reaction leads the researcher to consider much more carefully the methods and outcomes, and results in significantly improved efficiency and organisation.

Laboratory Archives and Repositories: The R4L Project

While the ELN is an ideal way to capture the experimental process and metadata, it may not be the most effective way to capture large quantities of experimental data. Spectra for example can consist of files many megabytes in size. The solution being developed in the Southampton "Repository for the Laboratory" R4L Project is to emulate the idea of laboratory chemical repositories and digital repositories for literature, to provide a location for the safe storage of laboratory data. This is similar to the repositories that have been successfully deployed for archiving and open access to papers. Access can be controlled and restricted to authorised users of the data. In due course the data can be made more widely available and linked to any publication that follows the investigation.

[5] ELNS: Worthy of note http://www.scientific-computing.com/scwjunjul06elns.html

Data Repository

The data files, are stored with a significant amount of metadata. Some are held in ASCII format but others are held in community-developed and internationally agreed formats. Where possible, different raw data are stored alongside any conversions (e.g. file formats or processed data) and "thumbnail" images for rapid identification. In the case of spectra, the key metadata item identifying the molecule (or at least the supposed molecular identity) is specified not only by the full chemical name but crucially the IUPAC InChI[6]. This will provide a digital URI capable of linking these resources with other chemical information. Sample specific URIs (i.e. a URI for this sample of the molecule made by a specific experiment) enable the links between the R4L entries and the ELN.

The R4L infrastructure is built upon the same eprints software used in the Southampton literature e-print system (also used for e-Crystals). This means that once the data are held within the R4L archive the metadata is searchable and can be exposed via the standard OAI systems. Metadata such as the molecular name, InChI, the type of spectra, etc, are exposed as additional information beyond the traditional Dublin Core, that carries the information on who created the data; the metadata are specified by a suitable schema.

The technical issues of the way R4L enables dissemination of the data are illustrated more fully by a comparable system, the e-Bank Project (see below the section Dissemination: The e-Bank Project), which actually preceded the R4L Project[7]. Related ideas are being developed in the Spectra Group[8] and the project web site provides details of how the results of some calculations are being archived on an institutionally focused subject repository.

Laboratory and Computational Data

An important aspect of the support for chemical investigations undertaken in the CombeChem Project, which is significant to the collaboration agenda, was the support of parallel experimental and theoretical work. One experimental project that was brought under the CombeChem work in this regard was the study of the behaviour of crown ether molecules at the air/water interface (Rousay, Fu, Robinson, Essex, & Frey, 2005). The experimental work involved a laser-based laboratory experiment while the theoretical work involved a large-scale molecular dynamics simulation. On the laboratory side, the work was used to trial the automatic capture of the laboratory context, i.e. the laboratory environmental conditions, the movements of people and similar information. On the simulation side, the fact that more information was derived from the joined-up study than from either of the individual studies alone was of interest to chemists.

The curation of laboratory data (consisting of many diverse but small datasets; methods, spectra, data from laser experiments, analysis and final fitted parameters), corresponded to the ideas discussed in previous sections. In contrast the requirements for curation of the simulation data were of a completely different magnitude. The

[6] International Union of Pure and Applied Chemistry (IUPAC) International Chemical Identifier (InChITM™) http://old.iupac.org/inchi/
[7] R4L: The Repository for the Laboratory - a JISC Project http://r4l.eprints.org/
[8] SPECTRa: Submission, Preservation and Exposure of Chemistry Teaching and Research Data http://www.lib.cam.ac.uk/spectra/

curation of something approaching a terabyte of simulation data represented a major concern. The solution being tested was an integration of the CombeChem Project with the software system provided by BioSimGrid e-Science Project[9] to hold and make available simulation results for further analysis simulation data. This proved a good test of BioSimGrid as the software was developed for bio-molecule simulation and not simulations containing a large number of separate molecules.

Initially the BioSimGrid system proved effective, holding the simulations and the necessary metadata and providing an analysis framework. However, in the longer term, the lack of a sustainability model for BioSimGrid is proving a difficulty, and we have had to store the simulation output in a more traditional manner. This is not ideal as while the recovery of the simulation output from the files is relatively straightforward, the necessary metadata needed to undertake a new analysis of the data are not readily available, and certainly not with the ease that BioSimGrid would provide this information.

The R4L (and e-Bank in general) model is for smaller-scale software solutions, individual research group or university-based repository solutions. While it still holds continuity problems for any individual dataset, it is more maintainable and sustainable than centralized software; even where, as in the case of BioSimGrid, the data were stored in a decentralized and replicated manner.

Data Dissemination: The e-Bank Project

The area of small molecule crystallography has many interesting aspects of potential use to a project aiming to extend the boundaries with respect to data curation in the chemical sciences. In many ways crystallography represents one of the most developed areas for curation of structural data. A global repository, the Cambridge Structural Database (CSD[10]) from Cambridge Crystallographic Data Centre (CCDC) has existed for some 30 years. Before publication of articles involving new crystal structures, all of the main chemical science journals make the following requirements: that the crystal structure encoded in the Crystallographic Information File format (CIF[11]) be deposited with, and validated by, the CCDC; then entered in the CSD. The CIF file contains the structural information in considerable detail in a format specified by the International Union of Crystallography (IUCr[12]), rendering it an ideal vehicle for the international dissemination of structural information obtained by crystallography.

The CIF format also builds on the fact that the workflow used in the analysis of crystallographic data is both well understood and agreed upon in the community; there are relatively few different programs being used. Although the CIF format is not XML, as the origin of the CIF format predates the main uptake of mark-up languages by scientific organizations, it does very carefully prescribe the necessary information in a strict format. The validation of the information presented in a CIF file, both in

[9] BioSimGrid: a distributed database for biomolecular simulations http://www.biosimgrid.org/
[10] The Cambridge Structural Database (CSD) http://www.ccdc.cam.ac.uk/products/csd/
[11] The International Union of Crystallography (IUCr) Crystallographic Information Framework http://www.iucr.org/iucr-top/cif/
[12] International Union of Crystallography (IUCr) http://www.iucr.org/

terms of completeness and that it accurately represents the results of the crystal structure determination, is a vital step in this process that adds value in a way that in the past was both very labour-intensive and time-consuming to achieve.

The situation would appear ideal: an internationally recognised format, a sustainable repository for the data, agreement by publishers and authors to use the database, and validation carried out by the database owners. Yet the current system is only delivering to the wider public some 20% of the structures that have been determined. The remaining data are being lost, are not really even informally curated, and are not accessible. This situation has not come about because of any deliberate action on the part of any of the players in the community; it arises from the need of researchers to concentrate on what they believe are the most important compounds. It takes too long to publish data that seem unimportant. Indeed a structure that proffers little or no novelty other than its not having been previously determined, can no longer be published through the usual mechanisms in a quality journal as it would fail to meet the usual publication criteria. Even for the data that are published, only the high-level structural information is accessible. The raw images, the stages of the analysis, are not available, even to reviewers of the publication. This represents a break in the provenance chain and prevents re-analysis of the data.

e-Bank and e-Crystals

The question dealt with by the e-Bank[13] team was how to use e-Science techniques to increase the exposure of the 80% of unpublished structures, while maintaining the necessary confidence in the disseminated information. Moreover an investigation was conducted to determine if such mechanisms could provide a richer source of information to the wider research and learning communities than is achieved by current methods (Coles et al., 2004; Heery et al., 2004). The work led to the creation of the e-Crystals repository and web site[14], which support the dissemination of crystal structures along with the raw and processed data that yielded the final CIF structure file. This much richer source of information is exposed via OAI[15] from a repository built on the Southampton e-print system[16] (Figure 4). At the highest (and simplest) level, this provides the molecular name, a short textual description of the project, the authors (both of the material and the structure determination), with the crystal structure (structural data) being designated an identifier (a DOI[17]). In this way it provides similar data to those lodged (for published structures with CCDC). It is very important to stress that the e-Crystals repository also furnishes the data produced by each of the major steps in the crystal structure determination workflow. It provides such files for all molecules entered into the repository, published or otherwise. These data and associated metadata describing these stages are characterised by a detailed schema, which serves to define rigorously the process represented in a crystal structure determination; and provide fine-grained provenance information sufficient for any citation of the structure, data or method.

[13] eBank UK Project http://www.ukoln.ac.uk/projects/ebank-uk/
[14] eCrystals - University of Southampton http://ecrystals.chem.soton.ac.uk/
[15] Open Archives Initiative http://www.openarchives.org/
[16] EPrints Free Software http://www.eprints.org/software/
[17] The Digital Object Identifier System http://www.doi.org/

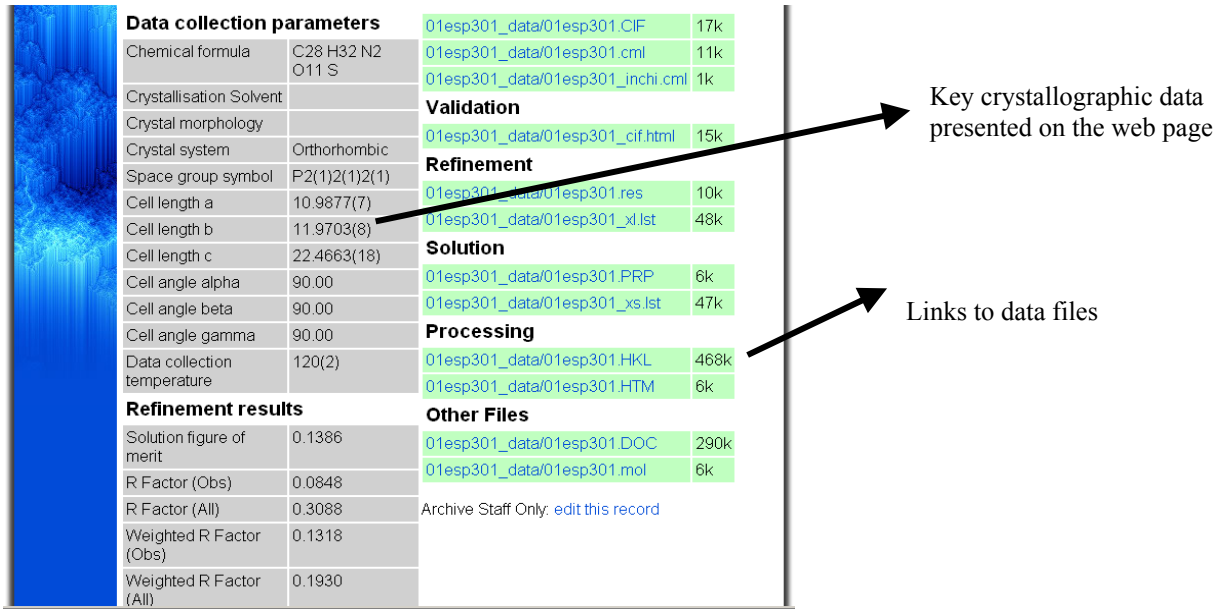

Figure 4. An example of the e-crystals repository showing the metadata provided about the crystal structure and the links to the data files for the provenance.

A user of the e-Crystals system (Coles et al., 2002, 2005, 2006; Coles, Frey, Peppe, et al., 2005), deposits the data files generated in a crystal structure solution in the e-print archive, which has been developed to integrate the crystal data using a schema based on extended Dublin Core[18]. A private laboratory version of e-Crystals is used as a laboratory management tool in the National Crystallographic Service. The data files are added as the analysis process is followed and then the end results are passed back to the users of the service. Under the agreement to use the service, if the structures are not published within three years, the structures will be made public via the institutional e-Crystals repository unless it is specially requested (with good grounds provided) that they remain private.

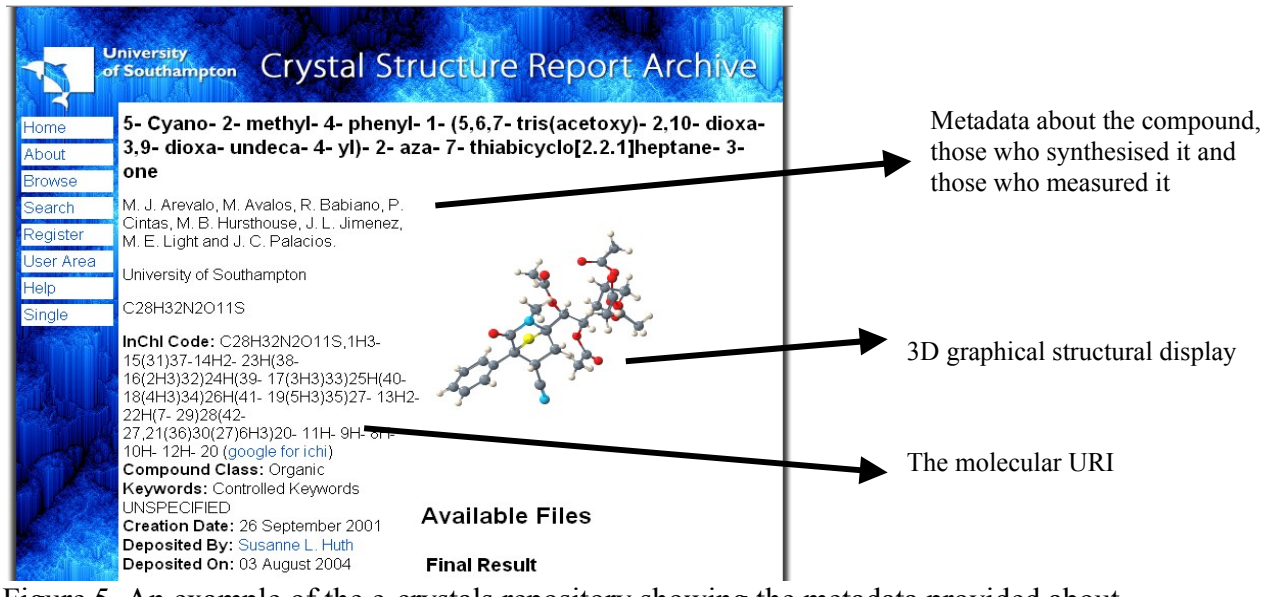

Figure 5. An example of the e-crystals repository showing the metadata provided about the crystal structure and the links to the data files for the provenance.

Using the e-Crystals approach, researchers can make their structural data available

[18] Dublin Core Metadata Initiative (DCMI) http://dublincore.org/

in a way that can be referenced and acknowledged. Data aggregators (such as CCDC) can "discover" the data and incorporate them into their databases and make them available to their audiences. The data are made available in a separate manner from any formal paper publication; indeed such a publication may not follow from the data, but other researchers may see the value of the structure especially as part of a wider group. The fact that making the data available in this way avoids peer review should not in this case be seen as a major problem. The full provenance of the data is available and exposed for general review in a way that few formally published structures are.

Validation

It is arguable whether it is worthwhile simply collating list after list of data, whether in print or on an electronic database, and expending resource on their maintenance. More valuable information results from a considered evaluation of the data, with recommendations of the best values and the reasons why these values have been selected. In the past, expert groups performed this type of evaluation. Thermodynamic data, for example, were evaluated by the National Institute for Science and Technology (NIST) (formerly the National Bureau of Standards (NBS)) and published as the JANAF tables (Gurvich, 1989). An IUPAC group reviews reaction kinetic data[19].

However, in many cases, the task of evaluation can no longer be assured through conventional mechanisms; essentially a vital step in the proper curation of the data is not being performed.

A partial solution to the lack of evaluation is to provide the users of the data with a much richer source of provenance information about the data. This approach will allow practitioners with at least some expertise in the area to make the evaluation themselves. In fact this wider community, having looked at the provenance and made use of the data, can then feed back views on their validity and accuracy. In effect this represents a wider community validation review - an ongoing curation activity. Furthermore, as the body of data stored with a full provenance trail increases, so does the possibility of much more automated checking of those data. In the area of thermodynamics most of the journals require authors to deposit their data in a NIST database[20]. Given the relationships that exist between thermodynamic quantities, it is possible to test the new data for consistency, highlight any potential problems and alert the authors. As the amount of data held in an easily convertible form increases, (i.e. XML with a proper schema so as to facilitate the comparison of related data from different sources), so this type of consistency checking, which provides a degree of automated validation, will become more widespread.

In a sense the e-Crystals Project is all about validation. Validation is provided in two ways. Firstly the validation process has been automated for some time, and the CIF files can be checked automatically for many errors in format and content (CIF Checker is provided by IUCr). This is one reason why small molecule crystallography is so ideal for this type of dissemination. Secondly, validation is also provided by the provision of the chain of analysis back to the raw data. It is possible therefore to verify

[19] IUPAC: Subcommittee for Gas Kinetic Data Evaluation http://www.iupac-kinetic.ch.cam.ac.uk/
[20] ThermoML: An XML-Based IUPAC Standard for Storage and Exchange of Experimental Thermophysical and Thermochemical Property Data http://trc.nist.gov/ThermoML.html

that the analysis has met particular requirements (though in instances of doubt the task of verification is often passed to an expert). While the relatively compact final and intermediate data files are held in the repository, the much larger raw data files (the service produces about 1Gb/day) are held securely at the Atlas Data Store[21] with a defined backup/recovery agreement. In principle a delocalised but linked system for multiple copies could also be employed.

Publication @ Source: R4L and the ELN Meet the Blog

In the CombeChem Project several of the components needed to improve the data trails between laboratory experiment and dissemination, and *vice versa*, have been investigated. The project has shown that it is possible to make use of planning information to: provide a context-aware digital capture right at the source of the investigations; capture the relationships between the processes and the data; support the archiving of these data; and provide a good basis for disseminating the data with all the relevant context. Some parts of the process are still missing; in particular the integration of all these steps has yet to reach the desired maturity.

The Semantic Blog

One methodology being investigated in the CombeChem Project is that of the "blog". Blogs have become a very versatile and fashionable means of capturing and facilitating a dialogue about events; an example is the blog run by *Scientific American*[22]. The capacity to link up different blogs and other web resources has made this a very powerful form of information sharing and integration.

The content of a laboratory notebook records a similar dialogue between researchers (largely one researcher's comments on a discussion that may involve many researchers) and experimental plans, procedures, data and analysis. The blog methodology appears to represent an excellent way in which to discuss and disseminate the information captured by the semantically aware ELNs. With research data placed in the blog or in data repositories, the experiments, data, analysis, and the scientific dialogue can be associated, and commented upon. At all times (i.e. all posts/comments) the links back to the raw data, other comments, and experimental description are easily maintained, once the data are in a blog. In this way the blog style captures all and more of the reports that would be present in the laboratory notebook. It is capable of capturing the links between discussions and allowing a discussion thread to be re-traced, far more readily than in a paper notebook, or indeed in the type of ELN that simply records in detail the results of an experiment, but without the ability to capture the context of that experiment.

Once the data have all been rationalized, the material in the blog provides all the information needed for a traditional publication. The publication should then refer back to the relevant blog posts, to provide the provenance for the ideas, and access to the experimental data, should others wish to use it (for the same or different purposes). Links work in both directions and while a post may be seen as a comment on some data, it provides at the same time additional metadata describing the data. So once relevant data have been located (publication -> blog -> repository) all comments made on those (or similar) data can be located, so that the full context of the data can be

[21] STFC e-Science Centre:Data Services Group
http://www.e-science.clrc.ac.uk/organisation/groups/data_services/
[22] *Scientific American* Observations Blog http://blog.sciam.com/

appreciated.

The blog approach to "open experimentation" is already being adopted by a few brave researchers and is providing a test of acceptability by both research staff and the wider scientific audience. A briefing on the UsefulChem malaria project (led by Jean-Claude Bradley) is given in the UsefulChem blog introduction[23] and contributions can be posted either in the discussion blog[24] or the molecule blog[25]. The blog of course lends itself to automatic and global exposure. A similar open report on a project developing the functionality of a blog is demonstrated by the authors' own work as part of a BBSRC-funded project, "Grid compatible data management for Directed Evolution Experiments", involving Cameron Neylon, Jeremy Frey and Jon Essex as well as students Jenny Hale and Andrew Milsted[26]. These projects are making a considerable mark in the quest to understand if Open Science can deliver the same transformative effects on development as Open Source has achieved in some aspects of computer software.

For adherents to a more traditional approach to science who wish to keep their data and results closer to home until time for publication, there is a need to limit the access to the blog to members of the research group, or formal collaborators. Security is however necessary in all cases, open or closed; to be useful as a proper laboratory record it is essential that all entries inspire confidence and are beyond unauthorised alteration. Perhaps more usefully, recognising that material does need to be updated and corrected, it would be wise to ensure that all alterations are clearly attributed. Keeping good records in this way within a blog means that this record can be made public as part of a publication.

Even though the blog function provides some of the necessary linkages between parts of the scientific discussion, adding links to data can be a barrier to maintaining a complete record. What is needed to ensure that the blog is both machine-readable and user-friendly, and able to provide the required links as automatically as possible, is a semantic blog. Semantic blog software represents a direct extension of the service-oriented approach as well as semantic underpinning of the ELNs; and as such is being built up under the CombeChem umbrella. A first stage in the process is to integrate the ideas from e-Crystals with the blog, using the full semantics and schema of e-Crystals to facilitate the relevant blog entries so that crystallographic material can be correctly placed within the blog automatically. We now have several pieces of equipment that blog their data; they have become "blog-jects". The data from the equipment automatically flow into the blog, for human comment and analysis. The environmental data of the experimental rooms are also recorded and blogged (a virtual blog) for later comparisons and correlations.

So far the discussion on the blog-notebooks has focused on their ability to complement the semantic ELN in recording and linking the scientific discussion. They thus act as a means to curate the scientific discourse. However the discussion would

[23] UsefulChem: malaria http://usefulchem.wikispaces.com/malaria
[24] Useful Chemistry http://usefulchem.blogspot.com/
[25] usefulchem-molecules http://usefulchem-molecules.blogspot.com/
[26] blogs@Chem*Tools*: The thoughts of Chemists: blog Beta-Glu
http://chemtools.chem.soton.ac.uk/projects/blog/blogs.php/blog_id/5

not be complete without some consideration of how the blog content should be curated itself. The Southampton blog system looks after the tracking of individual records via MD5 hashes of the entries to ensure that an unrecorded process has not changed the entry. Ultimately all the data are recorded in a database, which is backed up and archived. The basic content could easily be revived but would require the service to view all the inter-connections, which really provide the power of the system. We have at least transformed the problem from curating a wide scatter of seemingly unconnected items to that of curating the blog!

Conclusions

Those digital technologies that have dramatically accelerated the process of science have often rendered the curation of data in context more difficult. The metadata, the description of the process, become separated from the numerical data files, which, in principle, contain the data, but which, in practice, contain nothing useful without the contextual information. A better understanding and the improved support for semantics on the web mean that digital technology can now help to undo the damage that it has itself inflicted. Data curation can now be returned to those best placed to provide the context, namely the practitioners actually working to produce the data. They are the creators and disseminators of the material, and its quality is a direct result of their actions. Publication@source is already with us, even if we do not always recognise it. To facilitate the curation at the laboratory level, several projects (i.e. CombeChem, e-Bank, R4L, Spectra, BioSimGrid) are developing the necessary infrastructure to describe data, provide repositories to hold the data, and demonstrate how such data can be made available with appropriate provenance. With the right software in place, capturing the necessary metadata to ensure their provenance, and to facilitate subsequent curation, will become a natural consideration of the laboratory scientist.

Curation should be a matter of concern to the laboratory researcher. It should not be regarded as someone else's responsibility, nor undertaken at some late stage in the production of quality scientific data. Neither is curation simply a matter of ensuring that a scientific paper is kept available forever more. It is a matter of ensuring that quality data remain available and usable alongside the traditional paper. This process can be greatly facilitated by the actions of the researchers at the time the data are created and as such may require a significant shift in attitudes. Gathering metadata when they are available is much simpler and cheaper than trying to remember or re-construct them later.

If our data are important, we must design their curation into our experiments. In so doing we will not only develop better data, more useful to ourselves and to others, but also better science.

Acknowledgements

I would like to thank all of the CombeChem, e-Bank, and R4L research teams and in particular the support from my colleagues, L. Carr, S. J. Coles, D. C. DeRoure, J.W. Essex, M. B. Hursthouse, L. Lyon, D.C. Neylon, H.R. Mills, m. c. schraefel and students K. Taylor, J. Robinson, and A. Milsted and the EPSRC & JISC for funding. I would also like to thank reviewers for their most helpful feedback; I am especially grateful both to Chris Rusbridge and the Managing Editor whose suggestions have greatly improved this paper.

References

Allen, F.H. (2004). High-throughput crystallography: The challenge of publishing, storing and using the results. *Crystallography Reviews*, *10*, pp.3-15.

Atrium Research. (2006). Electronic laboratory notebook survey. Market Report A06-8. Retrieved June 26, 2008, from http://www.atriumresearch.com/html/eln_report2.htm

Butler, D. (2006). Electronic notebooks: A new leaf. *Nature*, *436*, 20.

Coles, S., Frey, J.G., Hursthouse, M.B., Light, M.E., Meacham, K.E., Marvin, D.J., & Surridge, M. (2005). ECSES - examining crystal structures using 'e-science': A demonstrator employing web and grid services to enhance user participation in crystallographic experiments. *Journal of Applied Crystallography*, *38, (5)*, 819-826. Retrieved June 13, 2008, from doi:10.1107/S0021889805025197.

Coles, S.J., Frey, J.G., Hursthouse, M.B., Light, M.E., Milsted, A.J., Carr, L.A., et al. (2006). An e-science environment for service crystallographys from submission to dissemination. *Journal of Chemical Information and Modeling*. Retrieved June 13, 2008, from doi:10.1021/ci050362w.

Coles, S.J., Frey, J.G., Hursthouse, M.B., Light, M.E., Surridge, M., Meacham, K.E., et al. (2002). Grid/Web enhancements to the National Crystallographic Service: Experiences with an interactive e-science demonstrator. In, F.R.A Hopgood, B. Matthews, & M.D. Wilson, (Eds.) *Euroweb 2002 - the Web and the GRID: from e-Science to e-Business , Oxford, UK, December 17-18, 2002*. Swindon, UK: British Computer Society. (Electronic Workshops in Computing). Retrieved June 30, 2008, from http://eprints.soton.ac.uk/346/

Coles, S, Frey, J.G., Peppe, S., Hursthouse, M., Light, M., Surridge, M., et al. (2005). Grid-enabling an existing instrument-based national service. In, H. Stockinger, R. Buyya, & R.Perrott, (Eds.) *Proceedings e-Science and Grid Computing 2005,* 570-577. (e-Science and Grid Computing, 1st International Meeting, Melbourne, Australia). Retrieved June 30, 2008, from http://eprints.soton.ac.uk/24099/

Coles, S., Lyon, L., Carr, L., Heery, R., Hursthouse, M., Gutteridge, C., et al. (2004). eBank UK: Linking research data, scholarly communication and learning. In *Semantic Grid Workshop, Global Grid Forum 11, Hawaii, USA, July 4-7, 2004*. Hawaii.

Dennis, C. (2002). Biology databases: Information overload. *Nature 417, 14*. Retrieved June 13, 2008, from doi: 10.1038/417014a

Frey, J.G., De Roure, D.C., & Carr, L. (2002). Publication at source: Scientific communication from a publication web to a data grid. Position Paper in *Euroweb 2002 the Web and the GRID: from e-Science to e-Business*. British Computer Society. Retrieved June 13, 2008, from http://ewic.bcs.org/conferences/2002/euroweb/session3/paper3.htm

Frey, J., De Roure, D., schraefel, m. c., Mills, H., Fu, H., Peppe, S., et al.. (2003). Context slicing the chemical aether. In M. David (Ed.), Proceedings of *First International Workshop on Hypermedia and the Semantic Web*. Nottingham, UK.

Frey, J.G., De Roure, D.C., Taylor, K., Essex, J., Mills, H., & Zaluska, E. (2006). CombeChem: A case study in provenance and annotation using the semantic web. *Provenance and Annotation of Data, International Provenance and Annotation Workshop, IPAW 2006*, 288. Retrieved June 13, 2008, from http://www.springer.com/uk/home/generic/search/results?SGWID=3-40109-22-173681711-0

Gurvich, L.V. (1989). Reference books and data banks on the thermodynamic properties of individual substances. *Pure & Applied Chemistry, 61*, 1027-1031.

Heery, R., Duke, M., Day, M., Lyon, L., Hursthouse, M. B., Frey, J.G., et al. (2004). Integrating research data into the publication workflow: The eBank UK experience. In *PV-2004: Ensuring the Long-Term Preservation and Adding Value to the Scientific and Technical Data, ESA/ESRIN, Frascati, Italy, 5 - 7 Oct 2004*. Frascati, Italy, European Space Agency, 8pp. Retrieved June 30, 2008, from http://www.ukoln.ac.uk/ukoln/staff/r.heery/publications.html#2004-10-07

Houlton, S. (2001). Data mining and the future of chemistry. *Manufact Chemist*, *72*, pp. 14-15.

Hoyt, D.W., Burton, S.D., Peterson, M.R., Myers, J.D., Chin, G. (2004). Expanding your laboratory by accessing collaboratory resources. *Anal and Bioanalytical Chemistry*, *378*, 1408-1410.

Hughes, G., Mills, H., de Roure, D., Frey, J., Moreau, L., schraefel, m. c., Smith, G., & Zaluska, E. (2004). The semantic smart laboratory: A system for supporting the chemical escientist. *Organic and Biomolecular Chemistry, 2*, pp. 1-10.

Myers, J.D., Allison, T.C., Bittner, S., Didier, B., Frenklach, M., Green, W.H., et al. (2005). A collaborative informatics infrastructure for multi-scale science. *Cluster Computing-The Journal of Networks Software Tools and Applications*, *8*, 243- 253.

Myers, J.D., Chappell, A.R., Elder, M., Geist, A., & Schwidder, J. (2003). Re-integrating the research record. *Computing in Science and Engineering*, *5*, 44-50.

Patterson, S.D. (2003). Data analysis—the Achilles heel of proteomics. *Nature Biotechnology*, *21*, pp.221-222. Retrieved June 13, 2008, from doi:10.1038/nbt0303-221

Rousay, E.R., Fu, H., Robinson, J.M., Essex, J.W., & Frey, J.G. (2005). Grid-based dynamic electronic publication: A case study using combined experiment and simulation studies of crown ethers at the air/water interface. *Philosophical Transactions A, 363*, 2075-2095. London: Royal Society.

Russo, E. (2002). Chemistry plans a structural overhaul. *Nature 419, 4-7*. Retrieved June 13, 2008, from doi: 10.1038/nj6903-04a

schraefel, m. c., Hughes, G., Mills, H., Smith, G., Payne, T., & Frey, J. (2004). Breaking the book: Translating the chemistry lab book into a pervasive computing lab environment. In Proceedings of *CHI 2004*. Vienna, Austria.